Numerical Optimisation: Conjugate gradient methods

Marta M. Betcke
m.betcke@ucl.ac.uk
Kiko Rullan
f.rullan@cs.ucl.ac.uk

Department of Computer Science
Centre for Medical Image Computing,
Centre for Inverse Problems
University College London

Lecture 5 & 6

- The linear CG method was proposed by Hestens and Stiefel in 1952 as a direct method for solution of linear systems of equations with positive definite matrix. (It was used to solve 106 difference equations on the Zuse computer at ETH (with a sufficiently accurate answer obtained in 90 iterations each approximately taking 2h 20 minutes.)
- In 1950 Lanczos iteration (including orthogonality of the basis and 3-term recurrence) applied to eigenvalue problems.
- Renaissance in early 1970 work by John Reid brought the connection to iterative methods. Game change: performance of CG is determined by the distribution of the eigenvalues of the matrix (preconditioning).
- In top 10 algorithms of 20th century.
- Nonlinear conjugate gradient method proposed by Fletcher and Reeves 1960.

Solution of linear system

$$Ax = b,$$

with A is symmetric positive definite matrix is equivalent to the quadratic optimisation problem

$$\min \phi(x) = \frac{1}{2}x^{\mathrm{T}}Ax - b^{\mathrm{T}}x.$$

Both have the same unique solution. In fact

$$\nabla\phi(x) = Ax - b =: r(x)$$

thus the linear system is the 1st order necessary condition (which is also sufficient for strictly convex function ϕ).

A set of non-zero vectors $\{p_1, p_2, \dots, p_\ell\}$ is said to be **conjugate** with respect to the symmetric positive definite matrix A if

$$p_i^{\mathrm{T}} A p_j = 0, \quad i \neq j, \quad i, j = 1, \dots, \ell.$$

Conjugate directions are linearly independent.

Conjugacy enables us to minimise ϕ in n steps by successfully minimising it along the individual directions in the set.

Given a starting point x_0 and the set of conjugate directions $\{p_0, p_2, \dots p_{n-1}\}$ let us generate the sequence $\{x_k\}$

$$x_{k+1} = x_k + \alpha p_k,$$

where α_k is the one dimensional minimiser of the quadratic function along p_k: $\phi(x_k + \alpha p_k)$

$$\alpha_k = -\frac{r_k^{\mathrm{T}} p_k}{p_k^{\mathrm{T}} A p_k}.$$

For any $x_0 \in \mathbb{R}^n$ the sequence converges to the solution $x^\star$ in at most n steps.

Proof: Because $\text{span}\{p_0, p_2, \dots p_{n-1}\} = \mathbb{R}^n$

$$x^\star - x_0 = \sigma_0 p_0 + \sigma_1 p_1 + \cdots + \sigma_{n-1} p_{n-1}.$$

Multiplying from the left by $p_k^{\mathrm{T}} A$ and using the conjugacy property we obtain σ_k as

$$\sigma_k = \frac{p_k^{\mathrm{T}} A (x^\star - x_0)}{p_k^{\mathrm{T}} A p_k}, \quad k = 0, \dots, n-1.$$

On the other hand, in the kth iteration the method generates approximation

$$x_k = x_0 + \alpha_0 p_0 + \alpha_1 p_1 + \cdots + \alpha_{k-1} p_{k-1}.$$

Multiplying from the left by $p_k^{\mathrm{T}} A$ and using the conjugacy property we have $p_k^{\mathrm{T}} A(x_k - x_0) = 0$ and

$$p_k^{\mathrm{T}} A(x^\star - x_0) = p_k^{\mathrm{T}} A(x^\star - x_k) = p_k^{\mathrm{T}} (b - Ax_k) = -p_k^{\mathrm{T}} r_k.$$

Substituting into $\sigma_k = -\frac{p_k^{\mathrm{T}} r_k}{p_k^{\mathrm{T}} A p_k} = \alpha_k$ for $k = 0, \dots, n-1$.

Theorem: [Expanding subspace minimisation]

For any starting point x_0 for the sequence $\{x_k\}$ generated by the conjugate direction method it holds

$$r_k^{\mathrm{T}} p_i = 0, \quad i = 0, 1, \dots, k-1,$$

and x_k is the minimiser of $\phi(a) = \frac{1}{2}x^{\mathrm{T}}Ax - b^{\mathrm{T}}x$ over the set

$$\{x \ : \ x = x_0 + \operatorname{span}\{p_0, p_1, \dots, p_{k-1}\}\}.$$

Proof: Let's define

$$h(\sigma) = \phi(x_0 + \sigma_0 p_0 + \cdots + \sigma_{k-1} p_{k-1}),$$

where $\sigma = (\sigma_0, \sigma_1, \ldots, \sigma_{k-1})^{\mathrm{T}}$. Since $h(\sigma)$ is a strictly convex quadratic, it has a unique minimiser $\sigma^\star$ that satisfies

$$\frac{\partial h(\sigma^\star)}{\partial \sigma_i} = 0, \quad i = 0, \ldots, k-1.$$

Using the chain rule we obtain

$$\nabla \phi(x_0 + \sigma^\star p_0 + \cdots + \sigma^\star_{k-1} p_{k-1})^{\mathrm{T}} p_i = 0, \quad i = 0, 1, \ldots, k-1.$$

Recall that $r(x) = \nabla\phi(x)$, thus for the minimiser $\tilde{x} = x_0 + \sigma^\star_0 p_0 + \cdots + \sigma^\star_{k-1} p_{k-1}$ on $\{p_0, p_1, \ldots, p_{k-1}\}$ it follows $r(\tilde{x})^{\mathrm{T}} p_i = 0$ as claimed.
By induction:
For $k = 1$, from $x_1 = x_0 + \alpha_0 p_0$ being a minimiser of ϕ along p_0 it follows $r_1^{\mathrm{T}} p_0 = 0$.

Suppose that $r_{k-1}^{\mathrm{T}} p_i = 0$ for $i = 0, 1, \dots, k-2$.

$$r_k = Ax_k - b = A(x_{k-1} + \alpha_{k-1} p_{k-1}) - b = r_{k-1} + \alpha_{k-1} A p_{k-1}.$$

and

$$p_{k-1}^{\mathrm{T}} r_k = p_{k-1}^{\mathrm{T}} r_{k-1} + \alpha_{k-1} p_{k-1}^{\mathrm{T}} A p_{k-1} = 0$$

by the definition of $\alpha_{k-1} = -\frac{p_{k-1}^{\mathrm{T}} r_{k-1}}{p_{k-1}^{\mathrm{T}} A p_{k-1}}$.
For any other $p_i, i = 0, 1, \dots, k-2$ we have

$$p_i^{\mathrm{T}} r_k = p_i^{\mathrm{T}} r_{k-1} + \alpha_{k-1} p_i^{\mathrm{T}} A p_{k-1} = 0,$$

where the first term disappears because of the induction hypothesis and the second because of the conjugacy of p_i. Thus we have shown $r_k^{\mathrm{T}} p_i = 0$ for $i = 0, 1, \dots, k-1$ and the proof is complete.

Conjugate gradient vs conjugate direction

- So far the discussion was valid for any set of conjugate direction.
- An example are eigenvectors of a symmetric positive definite matrix A which are orthogonal and conjugate w.r.t. A. Computation of full set of eigenvectors is expensive. Similarly, Gram Schmidt orthogonalisation process could be adopted to produce conjugate directions, however it is again expensive as it requires to store all the directions to orthogonalise against.
- Conjugate gradient (CG) method has a very special property, it can compute a new vector p_k using only the previous vector p_{k-1} i.e. it does not need to know the vectors $p_0, p_1, \dots p_{k-2}$ while p_k is automatically conjugate to those vectors. This makes CG particularly cheap in terms of computation and memory.

In CG each new direction is chosen as

$$p_k = -r_k + \beta_k p_{k-1},$$

where

$$\beta_k = \frac{r_k^{\mathrm{T}} A p_{k-1}}{p_{k-1}^{\mathrm{T}} A p_{k-1}}$$

follows from requiring that p_{k-1}, p_k be conjugate
i.e. $p_{k-1}^{\mathrm{T}} A p_k = 0$.

We initialise p_0 with the steepest descent direction at x_0.

As in the conjugate direction method, we perform successive one dimensional minimisation along each of the search directions.

CG: preliminary version

Given x_0

Set $r_0 = Ax_0 - b$, $p_0 = -r_0$, $k = 0$

while $r_k \neq 0$ **do**

$$\alpha_k = -\frac{r_k^{\mathrm{T}} p_k}{p_k^{\mathrm{T}} A p_k}$$

$$x_{k+1} = x_k + \alpha_k p_k$$

$$r_{k+1} = Ax_{k+1} - b$$

$$\beta_{k+1} = \frac{r_{k+1}^{\mathrm{T}} A p_k}{p_k^{\mathrm{T}} A p_k}$$

$$p_{k+1} = -r_{k+1} + \beta_{k+1} p_k$$

$$k = k + 1$$

end while

CG

Given x_0
Set $r_0 = Ax_0 - b$, $p_0 = -r_0$, $k = 0$
while $r_k \neq 0$ **do**

- $\alpha_k = \frac{r_k^{\mathrm{T}} r_k}{p_k^{\mathrm{T}} A p_k}$
- $x_{k+1} = x_k + \alpha_k p_k$
- $r_{k+1} = r_k + \alpha_k A p_k$
- $\beta_{k+1} = \frac{r_{k+1}^{\mathrm{T}} r_{k+1}}{r_k^{\mathrm{T}} r_k}$
- $p_{k+1} = -r_{k+1} + \beta_{k+1} p_k$
- $k = k + 1$

end while

Theorem:

For the kth iterate of the conjugate gradient method, $x_k \neq x^\star$ the following hold:

$$r_k^T r_i = 0, \quad i = 0, 1, \dots, k-1 \tag{1}$$

$$\mathrm{span}\{r_0, r_1, \dots, r_k\} = \underbrace{\mathrm{span}\{r_0, Ar_0, \dots, A^k r_0\}}_{=:\mathcal{K}_k(A, r_0)} \tag{2}$$

$$\mathrm{span}\{p_0, p_1, \dots, p_k\} = \mathrm{span}\{r_0, Ar_0, \dots, A^k r_0\} \tag{3}$$

$$p_k^T A p_i = 0, \quad i = 0, 1, \dots, k-1. \tag{4}$$

Therefore, the sequence $\{x_k\}$ converges to $x^\star$ in at most n steps.

The proof of this theorem relies on $p_0 = -r_0$. The result does not hold for other choices of p_0.

Note that the gradients r_k are actually orthogonal, while the directions p_k are conjugate, thus the name of conjugate gradients is actually a misnomer.

From the properties of the $k+1$st iterate we have

$$
\begin{aligned}
x_{k+1} &= x_0 + \alpha_0 p_0 + \dots + \alpha_k p_k && (5)\\
&= x_0 + \gamma_0 r_0 + \gamma_1 A r_0 + \dots + \gamma_k A^k r_0 && (6)
\end{aligned}
$$

for some γ_i, $i = 0, \dots, k$.
Let P_k denote the kth degree polynomial

$$P_k(\lambda) = \gamma_0 + \gamma_1 \lambda + \dots + \gamma_k \lambda^k$$

then

$$x_{k+1} = x_0 + P_k(A) r_0.$$

Recall that CG minimises the quadratic function ϕ over $x_0 + \operatorname{span}\{p_0, \dots, p_k\}$ which is the same as $x_0 + \mathcal{K}_k(A, r_0)$ i.e.

$$
\begin{aligned}
\arg\min \phi(x) &= \arg\min \phi(x) - \phi(x^\star)\\
&= \arg\min \frac{1}{2}(x - x^\star)^{\mathrm{T}} A (x - x^\star) = \arg\min \frac{1}{2}\|x - x^\star\|_A^2
\end{aligned}
$$

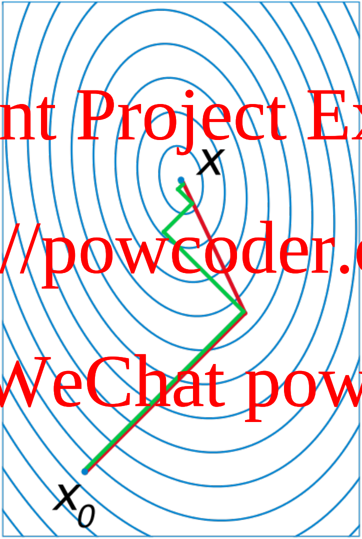

Figure: Wiki: Conjugate gradient method

Thus CG computes the minimising polynomial over all polynomials of degree k

$$\min_{P_k} \|x_0 + P_k(A) r_0 - x^\star\|_A.$$

Observe that similar expressions hold for the error

$$\begin{aligned} x_k - x^\star &= x_0 + P_{k-1}(A) r_0 - x^\star = x_0 - x^\star + P_{k-1}(A) \underbrace{A(x_0 - x^\star)}_{=r_0} \\ &= [I + AP_{k-1}(A)](x_0 - x^\star) \end{aligned}$$

and the residual

$$\begin{aligned} r_k &= Ax_k - b = A(x_k - x^\star) = A(x_0 + P_{k-1}(A) r_0 - x^\star) \\ &= \underbrace{A(x_0 - x^\star)}_{=r_0} + AP_{k-1}(A) r_0 = [I + AP_{k-1}(A)] r_0 \end{aligned}$$

Let the eigenvalue decomposition of the symmetric positive definite matrix

$$A = V^{\mathrm{T}} \Lambda V = \sum_{i}^{n} \lambda_i v_i v_i^{\mathrm{T}},$$

with $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ and $v_i, i = 1, \ldots, n$ the corresponding orthogonal eigenvectors.

Since V is a basis for any vector in $\mathbb{R}^n$, in particular for $x_0 - x^\star = \sum_{i=1}^n \xi_i v_i$.

Notice that any eigenvector v_i of A is also an eigenvector of $P_k(A)$ with the corresponding eigenvalue $P_k(\lambda_i)$

$$P_k(A) v_i = P_k(\lambda_i) v_i, \quad i = 1, \ldots, n.$$

Hence

$$x_{k+1} - x^\star = \sum_{i=1}^{n} [1 + \lambda_i P_k(\lambda_i)] \xi_i v_i$$

and

$$\|x_{k+1} - x^\star\|_A^2 = \sum_{i=1}^{n} \lambda_i [1 + \lambda_i P_k(\lambda_i)]^2 \xi_i^2$$

Since P_k is optimal w.r.t. this norm we have

$$\begin{aligned}
\|x_{k+1} - x^\star\|_A^2 &= \min_{P_k} \sum_{i=1}^{n} \lambda_i [1 + \lambda_i P_k(\lambda_i)]^2 \xi_i^2 \\
&\le \min_{P_k} \max_{1 \le i \le n} [1 + \lambda_i P_k(\lambda_i)]^2 \left(\sum_{i=1}^{n} \lambda_i \xi_i^2 \right) \\
&= \min_{P_k} \max_{1 \le i \le n} [1 + \lambda_i P_k(\lambda_i)]^2 \|x_0 - x^\star\|_A^2.
\end{aligned}$$

Theorem If A has only r distinct eigenvalues, then CG will converge to the solution in at most r iterations.

Proof: Suppose the eigenvalues take on distinct r values $\tau_1 < \dots < \tau_r$ and define a polynomial

$$Q_r(\lambda) = \frac{(-1)^r}{\tau_1 \tau_2 \dots \tau_r}(\lambda - \tau_1)\dots(\lambda - \tau_r)$$

and note that $Q_r(\lambda_i) = 0$, $i = 1, \dots, n$ and $Q(0) = 1$. Then

$$\bar{P}_{r-1}(\lambda) = (Q_r(\lambda) - 1)/\lambda$$

is of degree $r-1$ and we have

$$\begin{aligned} 0 &\leq \min_{P_{r-1}} \max_{1 \leq i \leq n} [1 + \lambda_i P_{r-1}(\lambda_i)]^2 \\ &\leq \max_{1 \leq i \leq n} [1 + \lambda_i \bar{P}_{r-1}(\lambda_i)]^2 = \max_{1 \leq i \leq n} Q_r^2(\lambda_i) = 0 \end{aligned}$$

and $\|x_r - x^\star\|_A^2 = 0$ and hence $x_r = x^\star$.

Theorem If A has eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$, we have that

$$\|x_{k+1} - x^\star\|_A^2 \leq \left(\frac{\lambda_{n-k} - \lambda_1}{\lambda_{n-k} + \lambda_1} \right)^2 \|x_0 - x^\star\|_A^2.$$

Proof idea: Choose polynomial $\bar{P}_k$ such that $Q_{k+1}(\lambda) = 1 + \lambda \bar{P}_k(\lambda)$ has roots at the k largest eigenvalues $\lambda_n, \lambda_{n-1}, \dots, \lambda_{n-k+1}$ and at the midpoint between λ_{n-k} and λ_1. It can be shown that the maximum value attained by Q_{k+1} on the remaining eigenvalues $\lambda_1, \dots, \lambda_{n-k}$ is $\frac{\lambda_{n-k} - \lambda_1}{\lambda_{n-k} + \lambda_1}$.

Theorem In terms of condition number $\kappa(A) = \|A\|_2 \|A^{-1}\|_2 = \lambda_n / \lambda_1$, we have that

$$\|x_k - x^\star\|_A \leq 2 \left(\frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1} \right)^k \|x_0 - x^\star\|_A.$$

Preconditioning

We can accelerate CG through transformations which cluster eigenvalues. This process is known as **preconditioning**.

We perform a change of variables $\hat{x} = Cx$.

Then the quadratic function ϕ in terms of $\hat{x}$ reads

$$\hat{\phi}(\hat{x}) = \frac{1}{2}\hat{x}^{\mathrm{T}}(C^{-\mathrm{T}}AC^{-1})\hat{x} - (C^{-\mathrm{T}}b)^{\mathrm{T}}\hat{x}.$$

Minimising $\hat{\phi}$ is equivalent to solving the system of normal equations

$$(C^{-\mathrm{T}}AC^{-1})\hat{x} = (C^{-\mathrm{T}}b)$$

and the convergence rate of CG depends on the eigenvalues of $C^{-\mathrm{T}}AC^{-1}$.

It is not necessary to carry out the transforms explicitly. We can apply CG to $\hat{\phi}$ in terms of $\hat{x}$ and then invert the transformations to reexpress all the equations in terms of the original variable x.

In fact, the **preconditioned CG** algorithm does not use the factorisation $M = C^{\mathrm{T}}C$ explicitly, only M.

If we set $M = I$ we recover unpreconditioned CG algorithm.

The properties of CG generalise, in particular for PCG it holds

$$r_i^{\mathrm{T}} M^{-1} r_j = 0, \quad \forall i \neq j.$$

Preconditioned CG (PCG)

Given x_0, preconditioner M

Set $r_0 = Ax_0 - b$,

Solve $My_0 = r_0$

$p_0 = -y_0$, $k = 0$

while $r_k \neq 0$ **do**

$\alpha_k = \frac{r_k^{\mathrm{T}} y_k}{p_k^{\mathrm{T}} A p_k}$

$x_{k+1} = x_k + \alpha_k p_k$

$r_{k+1} = r_k + \alpha_k A p_k$

Solve $My_{k+1} = r_{k+1}$

$\beta_{k+1} = \frac{r_{k+1}^{\mathrm{T}} y_{k+1}}{r_k^{\mathrm{T}} y_k}$

$p_{k+1} = -y_{k+1} + \beta_{k+1} p_k$

$k = k + 1$

end while

Nonlinear conjugate gradient

Recall that CG can be interpreted as a minimiser of a quadratic convex function

$$\phi(x) = \frac{1}{2}x^{\mathrm{T}}Ax - x^{\mathrm{T}}b$$

Can the algorithm for ϕ be generalised to a nonlinear function f?

Recall that

- step length α_k minimises ϕ along p_k.
  For general f: compute α_k using line search

$$\alpha_k = \min_{\alpha} f(x_k + \alpha p_k)$$

- $r = Ax - b = \nabla\phi(x)$.
  For general function f: $r \to \nabla f$

Given x_0

Evaluate $f_0 = f(x_0)$, $\nabla f_0 = \nabla f(x_0)$

Set $p_0 = -\nabla f_0$, $k = 0$

while $\nabla f_k \neq 0$ **do**

Compute α_k using line search: $\alpha_k = \min_\alpha f(x_k + \alpha p_k)$

$x_{k+1} = x_k + \alpha_k p_k$

$\beta_{k+1} = \frac{\nabla f_{k+1}^{\mathrm{T}} \nabla f_{k+1}}{\nabla f_k^{\mathrm{T}} \nabla f_k}$

$p_{k+1} = -\nabla f_{k+1} + \beta_{k+1} p_k$

$k = k + 1$

end while

Is p_k a descent direction?

$$\nabla f_k^{\mathrm{T}} p_k = -\nabla f_k^{\mathrm{T}} \nabla f_k + \beta_k \nabla f_k^{\mathrm{T}} p_{k-1} \overset{?}{<} 0$$

If α_{k-1} is a local minimiser along p_{k-1}, $\nabla f_k^{\mathrm{T}} p_{k-1} = 0$ and

$$\nabla f_k^{\mathrm{T}} p_k = -\nabla f_k^{\mathrm{T}} \nabla f_k < 0$$

thus p_k is a descent direction.

If the linear search is not exact, due to the second term $\beta_k \nabla f_k^{\mathrm{T}} p_{k-1}$, p_k may fail to be a descent direction. This can be avoided by requiring that the step length α_{k-1} satisfies the strong Wolfe conditions

$$\begin{aligned} f(x_k + \alpha_k p_k) &\leq f(x_k) + c_1 \alpha_k \nabla f_k^{\mathrm{T}} p_k, \\ |\nabla f(x_k + \alpha_k p_k)^{\mathrm{T}} p_k| &\leq -c_2 \nabla f_k^{\mathrm{T}} p_k \end{aligned}$$

with $0 < c_1 < c_2 < \frac{1}{2}$.

Let f be twice continuously differentiable, and the level set $\{x : f(x) \leq f(x_0)\}$ is bounded. If the step length α_k in the FR algorithm satisfies strong Wolfe conditions with $0 < c_2 < \frac{1}{2}$, then the method generates descent directions p_k that satisfy

$$-\frac{1}{1-c_2} \leq \frac{\nabla f_k^{\mathrm{T}} p_k}{\|\nabla f_k\|^2} \leq \frac{2c_2 - 1}{1-c_2}, \quad k = 1, 2, \dots \tag{7}$$

Proof: First note that the upper bound $(2c_2 - 1)/(1 - c_2)$ monotonically increases for $c_2 \in (0, \frac{1}{2})$ and $-1 < (2c_2 - 1)/(1 - c_2) < 0$. Thus Lemma [1] implies that p_k is a descent direction $\nabla f_k^{\mathrm{T}} p_k < 0$.
The inequalities can be shown by induction using the form of the update the second strong Wolfe condition.

Induction:

$k = 0: \quad p_0 = -\nabla f_0 \to \frac{\nabla f_0^{\mathrm{T}} p_0}{\|\nabla f_0\|^2} = -1$ and (7) holds.

Assume (7) holds for some $k \geq 1$. From β_{k+1}^{FR} we have

$$\frac{\nabla f_{k+1}^{\mathrm{T}} p_{k+1}}{\|\nabla f_{k+1}\|^2} = -1 + \beta_{k+1}^{FR} \frac{\nabla f_{k+1}^{\mathrm{T}} p_k}{\|\nabla f_{k+1}\|^2} = -1 + \frac{\nabla f_{k+1}^{\mathrm{T}} p_k}{\|\nabla f_k\|^2}$$

Plugging curvature Wolfe condition $|\nabla f_{k+1}^{\mathrm{T}} p_k| \leq -c_2 \nabla f_k^{\mathrm{T}} p_k$ into last equation (note $\nabla f_k^{\mathrm{T}} p_k < 0$ by induction hypothesis) we obtain

$$-1 + c_2 \frac{\nabla f_k^{\mathrm{T}} p_k}{\|\nabla f_k\|^2} \leq \frac{\nabla f_{k+1}^{\mathrm{T}} p_{k+1}}{\|\nabla f_{k+1}\|^2} \leq -1 - c_2 \frac{\nabla f_k^{\mathrm{T}} p_k}{\|\nabla f_k\|^2}.$$

Substituting the lower bound for $\frac{\nabla f_k^{\mathrm{T}} p_k}{\|\nabla f_k\|^2}$ from induction hypothesis we obtain (7) for $k+1$

$$-1 - \frac{c_2}{1-c_2} \leq \frac{\nabla f_{k+1}^{\mathrm{T}} p_{k+1}}{\|\nabla f_{k+1}\|^2} \leq -1 + \frac{c_2}{1-c_2}.$$

Weakness of FR algorithm

If FR generates a bad direction and a tiny step, then the next direction and the next step are also likely to be poor.

Let $\theta_k = \angle(p_k, -\nabla f_k)$,

$$\cos\theta_k = \frac{-\nabla f_k^T p_k}{\|\nabla f_k\| \|p_k\|}.$$

A bad direction p_k is almost orthogonal to $-\nabla f_k$ and $\cos\theta_k \approx 0$.

Multiplying (7) by $\|\nabla f_k\| / \|p_k\|$ we obtain

$$\frac{1-2c_2}{1-c_2} \frac{\|\nabla f_k\|}{\|p_k\|} \leq \cos\theta_k \leq \frac{1}{1-c_2} \frac{\|\nabla f_k\|}{\|p_k\|}, \quad k = 1, 2, \ldots.$$

Thus $\cos\theta_k \approx 0$ if and only if $\|\nabla f_k\| \ll \|p_k\|$.

Since p_k is almost orthogonal to $-\nabla f_k$, the step from x_k to x_{k+1} is likely tiny, i.e. $x_{k+1} \approx x_k$. Consequently, $\nabla f_k \approx \nabla f_{k+1}$ then $\beta_{k+1} \approx 1$ and finally given $\|\nabla f_{k+1}\| \approx \|\nabla f_k\| \ll \|p_k\|$, $p_{k+1} \approx p_k$ and the new direction will improve little.

If $\cos\theta_k \approx 0$ holds and the the subsequent step is small, the following updates are unproductive.

Polak-Ribière:

$$\beta_{k+1}^{PR} = \frac{\nabla f_{k+1}^{\mathrm{T}}(\nabla f_{k+1} - \nabla f_k)}{\|\nabla f_k\|^2} \tag{8}$$

If f is strongly convex quadratic function and the line search is exact, $\nabla f_{k+1} \perp \nabla f_k$ and $\beta_{k+1}^{PR} = \beta_{k+1}^{FR}$.

For general nonlinear functions and inexact line search, numerical experience indicates that PR algorithm is more robust and efficient.

As is, the strong Wolfe conditions do not guarantee that p_k is always a descent direction. For $\beta_{k+1} = \max\{\beta_{k+1}^{PR}, 0\}$, simple adaptation of strong Wolfe conditions ensures the descent property.

Hestenes - Stiefel (similar to PR in both theory and practical performance):

Consecutive directions are conjugate wrt *average Hessian* $\bar{G}_k = \int_0^1 \nabla^2 f(x_k + \tau \alpha_k p_k) d\tau$. From Taylor's theorem we have $\nabla f_{k+1} = \nabla f_k + \alpha_k \bar{G}_k p_k$. Solving $p_{k+1}^{\mathrm{T}} \bar{G}_k p_k = 0$ where $p_{k+1} = -\nabla f_{k+1} + \beta_{k+1} p_k$ for β_{k+1} yields

$$\beta_{k+1}^{HS} = \frac{\nabla f_{k+1}^{\mathrm{T}} (\nabla f_{k+1} - \nabla f_k)}{(\nabla f_{k+1} - \nabla f_k)^{\mathrm{T}} p_k} \qquad (9)$$

Two competitive with PR choices which guarantee p_k to be descent direction under (standard) Wolfe conditions on α_k:

$$\beta_{k+1} = \frac{\|\nabla f_{k+1}\|^2}{(\nabla f_{k+1} - \nabla f_k)^{\mathrm{T}} p_k} \qquad (10)$$

$$\beta_{k+1} = \left(y_k - 2 p_k \frac{\|y_k\|^2}{y_k^{\mathrm{T}} p_k} \right)^{\mathrm{T}} \frac{\nabla f_{k+1}}{y_k^{\mathrm{T}} p_k} \quad \text{with } y_k = \nabla f_{k+1} - \nabla f_k. \qquad (11)$$

Set $\beta_k = 0$ in every nth step i.e. take steepest descent step. Restarting serves to refresh the algorithm erasing old information that may be not beneficial. Such restarting leads to n step quadratic convergence $\|x_{k+n} - x^\star\| = \mathcal{O}\|x_k - x^\star\|^2$.

Consider function which is strongly convex quadratic close to the solution $x^\star$ but non-quadratic elsewhere. Once close to the solution the restart will allow the method to behave like linear conjugate gradients, in particular with finite termination within n steps from the restart (recall that the finite termination property for linear CG only holds if initiated with $p_0 = -\nabla f_0$).

In practice, conjugate gradient methods are usually used when n is large, hence n steps are never taken. Observe that the gradients are mutually orthogonal when f is a quadratic function. Restart when two consecutive gradients are far from orthogonal $\frac{|\nabla f_k^{\mathrm{T}} \nabla f_{k+1}|}{\|\nabla f_k\|^2} \geq \nu$, with ν typically 0.1.

When for some search direction p_k, $\cos\theta_k \approx 0$ and the subsequent step is small, substituting $\nabla f_{k+1} \approx \nabla f_k$ into β_{k+1}^{PR} results in $\beta_{k+1}^{PR} \approx 0$ and the next direction $p_{k+1} \approx -\nabla f_{k+1}$ the steepest descent direction. Therefore the PR algorithm essentially performs a restart after it encounters a bad direction.

The same argument applies to HS, and PR+.

FR algorithm requires some restart.

Hybrid FR-PR:
Global convergence can be guaranteed if $|\beta_k| \leq \beta_k^{FR}$ for all $k \geq 2$. This suggest following strategy:

$$\beta_k = \begin{cases} -\beta_k^{FR}, & \beta_k^{PR} < -\beta_k^{FR} \\ \beta_k^{PR}, & |\beta_k^{PR}| \leq \beta_k^{FR} \\ \beta_k^{FR}, & \beta_k^{PR} > \beta_k^{FR} \end{cases} \tag{12}$$

Global convergence - assumptions

Assumptions:

i) The level set $\mathcal{L} = \{x : f(x) \leq f(x_0)\}$ be bounded.

ii) In some open neighbourhood $\mathcal{N}$ of $\mathcal{L}$, the objective function f is Lipschitz continuously differentiable.

These assumptions imply that there is a constant γ such that

$$\|\nabla f(x)\| \leq \gamma, \forall x \in \mathcal{L}.$$

From Zoutenjik's lemma it follows that any line search iteration $x_{k+1} = x_k + \alpha_k p_k$ where p_k is a descent direction and the step length α_k satisfies Wolfe conditions gives the limit

$$\sum_{k=0}^{\infty} \cos^2 \theta_k \|\nabla f_k\|^2 < \infty.$$

Similarly, to the global convergence for line search, global convergence for **restarted** conjugate gradient algorithms periodically setting $\beta_k = 0$ (hence $\cos \theta_{k+1} = 1$) can be proven in a subsequence

$$\lim_{k \to \infty} \inf \|\nabla f_k\| = 0.$$

Theorem: [Al-Baali] Suppose that the assumptions i) and ii) hold and FR algorithm is implemented with line search that satisfies strong Wolfe conditions with $0 < c_1 < c_2 < \frac{1}{2}$. Then

$$\lim_{k\to\infty} \inf \|\nabla f_k\| = 0.$$

Proof: By contradiction (assume $\|\nabla f_k\| \geq \gamma > 0$). Substitute into Zoutenjik's result and use definition of p_k and upper bound in Lemma [1] recursively to show that the assumed to converge sequence in lower bounded by harmonic series which is divergent hence contradiction.

This global convergence result can be extended to any method satisfying $|\beta_k| \leq \beta_k^{FR}$ for all $k \geq 2$.

If constants $c_4, c_5 > 0$ exist such that

$$\cos\theta_k \geq c_4 \frac{\|\nabla f_k\|}{\|p_k\|}, \quad \frac{\|\nabla f_k\|}{\|p_k\|} \geq c_5 > 0, \quad k = 1, 2, \dots$$

if follows from Zoutenjik's result that

$$\lim_{k\to\infty} \|\nabla f_k\| = 0.$$

This result can be established for PR for f strongly convex and exact line search.

For general nonconvex functions it is not possible even though PR performs better in practice than FR. PR method can cycle infinitely even if ideal line search is used i.e. line search which returns α_k that is the first positive stationary point of $f(x_k + \alpha p_k)$. Example relies on $\beta_k < 0$ which motivated the modification $\beta_k^+ = \max\{0, \beta_k\}$.